IMPACT OF DEMONSTRATION VARIABILITY ON DEEP NEURAL NETWORK PERFORMANCE IN SIGN LANGUAGE RECOGNITION

[1]NURZADA AMANGELDY, [2]BEKBOLAT KURMETBEK, [3]GAZIZOVA NAZERKE, [4]SAULE KUDUBAYEVA, [5]RAUSHAN AMANGELDY

[1,2,3,4,5] L.N.Gumilyov Eurasian National University, 2, Satpayev str., Astana, 010000, Republic of Kazakhstan
Email: [1]nurzadaamangeldy@gmail.com

Abstract - This article investigates the impact of demonstration variability on the accuracy of recognizing sign language words and alphabet using deep neural networks. The primary focus is on evaluating the effectiveness of various deep learning architectures. The key contribution of this work lies in identifying the importance of multimodal approaches and model adaptation to enhance the accuracy of gesture recognition under conditions of low variability, specifically using the alphabet as an example. This is crucial for the development of automatic sign language and alphabet recognition systems. It is particularly critical because, unlike dynamic words, proper nouns demonstrated through the alphabet must be recognized with consistent efficiency during continuous recognition.

Keywords - gesture recognition, variability, lip reading, multimodal approaches, clustering, model adaptation, real-time recognition

I. INTRODUCTION

Sign Language and Alphabet Recognition Tasks Have Advanced Significantly Thanks to the Integration of Multimodal Deep Learning Methods. Researchers have proposed various structures and architectures to enhance the accuracy and reliability of gesture recognition systems. Let's examine several solutions, grouped by the similarity of the deep learning models used.

The following approaches demonstrate the effectiveness of recurrent neural networks in gesture analysis under various conditions. The DeepArSLR model includes hand segmentation and deep recurrent neural networks (LSTM), making it suitable for recognizing sign language words where body movements and poses are important. This approach is well-suited for tasks requiring the analysis of temporal gesture sequences [1].The SIGNFORMER model combines Pose LSTM with Pose Transformer, utilizing the self-attention mechanism, allowing effective recognition of both sign language words and the sign alphabet [2].

Convolutional Neural Networks (CNNs) have shown high performance in gesture recognition tasks. A real-time CNN model tailored for gesture recognition has demonstrated good results with a limited number of images, making it suitable for tasks requiring quick and accurate recognition of alphabet gestures [3].A hybrid model based on VGG16 and an ensemble classifier optimized for gesture recognition has also shown excellent results in terms of execution time and accuracy [4].

Vision Transformer (ViT) application for static gesture recognition has achieved high accuracy. For the sign alphabet, where hand shape recognition is crucial, this model can be particularly effective [5].

Generative Adversarial Networks (GANs) provide a context-dependent approach for continuous gesture recognition, aiding in recognizing a continuous stream of gestures [6]. A graph convolutional neural network with spatial attention improves gesture recognition accuracy and is especially useful for analyzing whole-body movements [7].The PG-GCN model uses pose and skeleton data for action recognition, enhancing the accuracy of recognizing both sign language words and the alphabet [8].

Multimodal integrations improve gesture recognition models' capabilities. The late fusion multimodal approach in gesture recognition outperforms single-modal approaches, which is particularly beneficial for recognizing sign language words [9]. A multiscale and dual graph convolutional network demonstrates its effectiveness in gesture recognition, matching the results of state-of-the-art systems [10].A model using singular value entropy and deep belief networks also enhances gesture recognition, showing high performance [11].

Different deep learning models can be used to solve the tasks of sign language and sign alphabet recognition. Recurrent neural networks and transformers are suitable for tasks requiring the analysis of temporal sequences and spatial data. Convolutional neural networks are effective for recognizing hand shapes in the sign alphabet. Generative adversarial networks and graph convolutional networks improve the recognition of complex movements and contexts. The use of multimodal data and various architectures significantly enhances the accuracy and reliability of gesture recognition systems.

Based on the above studies and approaches, several key questions arise that require further investigation to improve gesture recognition systems. These questions will help deepen the understanding of the problem and find ways to enhance the accuracy and reliability of deep learning models, especially considering their practical applications.

Research Questions for Experimentation

1. Is it possible to use a single architecture for recognizing both the sign language alphabet and sign language words?

This question aims to explore the feasibility of applying a unified deep learning model for the recognition of different types of gestures. How effectively can the same architecture handle the tasks of recognizing static alphabet gestures as well as more variable dynamic word gestures? What modifications or adaptations to the model might be necessary to achieve high performance in both cases?

2. How to develop a solution that significantly improves the recognition of the sign language alphabet while efficiently recognizing sign language words as an addition?

This question focuses on investigating ways to enhance the accuracy of sign language alphabet recognition while maintaining the efficiency of sign language word recognition. What data preprocessing methods, algorithms, or hybrid models can be employed to create a system that excels at recognizing both the alphabet and words? What architectural designs and multimodal approaches can be integrated to improve the overall performance of the gesture recognition system?

II. DETAILS EXPERIMENTAL

2.1 Data collecting

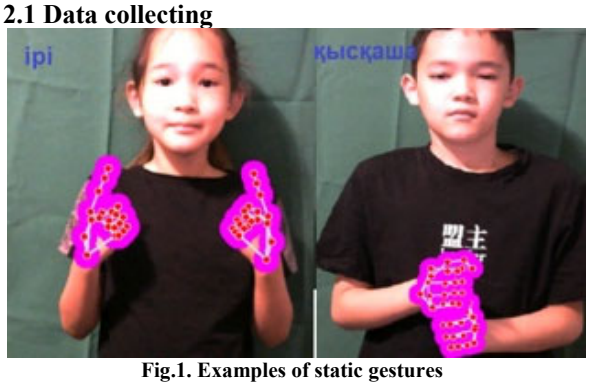


Fig.1. Examples of static gestures

To create and train the model for recognizing the Kazakh dactyl alphabet, consisting of 42 letters, video materials were collected. Three demonstrators participated in the experiment: two professional sign language interpreters and one project staff member. Each demonstrator was recorded twice: once with a focus on the clarity and precision of gestures, and a second time including pronunciation and lip movements, thus creating a comprehensive database. As a result, 60 video recordings were obtained for each letter of the 42 letters in the Kazakh dactyl alphabet, totaling 2520 video recordings. This database was designed to provide a comprehensive approach to training the model, allowing for the consideration of both visual gestures and lip articulation.

2.2 Architecture and Training Process

For training the model, LSTM1024 and 2D CNN architectures [12, 13] were used, specifically adapted for gesture classification tasks. In previous works focused on the recognition of sign language words using the same deep neural networks, the models demonstrated high accuracy.

In previous experiments on sign language word recognition using the 2D CNN architecture, the model achieved high accuracy in gesture classification. Other contemporary video datasets, such as AUTSL and LSA64, were trained using the proposed architecture and showed good results. In these experiments, forty sign language words were selected based on maximum movement variability, allowing the model to effectively handle various forms of gesture demonstration. Specifically, for the classification of Kazakh Sign Language (KSL) gestures, the 2D CNN architecture, in combination with other methods such as computer vision techniques and motion capture algorithms, demonstrated good results. These experiments confirmed that the use of multilayer perceptrons and convolutional neural networks allows achieving high accuracy and reliability in gesture recognition.

However, applying the same architecture to recognize the Kazakh dactyl alphabet yielded less impressive results. The main problem was the decrease in accuracy on validation datasets, indicating the need for further tuning and improvement of the model. This is due to the fact that the alphabet requires more detailed processing of visual data and less variability in gestures compared to sign language words.

The variability of sign language words significantly affects the efficiency of gesture recognition, as gestures can vary greatly in terms of execution location,shape, movement direction, and other parameters [14, 15]. Depending on these factors, the recognition task can become significantly more complex. For instance, if a gesture is performed in a neutral zone, the object can be easily separated from the background using the speaker's solid-colored clothing. However, if a gesture is performed near the face or neck, the hand may blend with the face or neck, complicating gesture recognition. For gestures performed at waist or shoulder level, certain background requirements are necessary to ensure accurate object recognition.

Gestures can be either static or dynamic (Fig. 1,2,3). Static gestures represent a hand position without movement in space, whereas dynamic gestures are characterized by successive hand movements from an initial to a final point over a certain period of time. Examples of static gestures include letters of the dactyl alphabet, which often represent fixed hand positions.

Additionally, variability is evident in the use of one or both hands when performing gestures. Symmetrical two-handed gestures are characterized by the matching shape and direction of movement of both hands.



Fig.2.Symmetrical two-handed gestures

In asymmetrical two-handed gestures, one hand often remains stationary or moves in a different direction, serving as the passive hand, while the other hand performs complex movements, acting as the active hand.

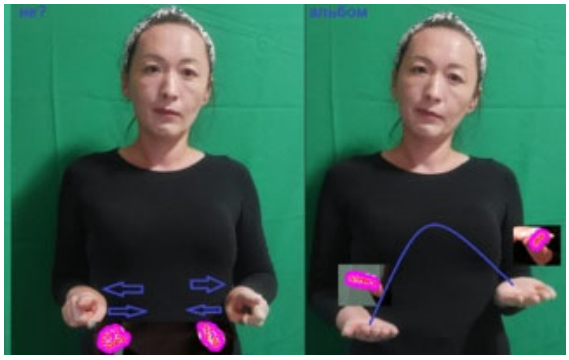


Fig.3.Asymmetrical two-handed gestures

Such characteristics impose additional requirements on recognition algorithms, which must take into account the configuration and movement of the active hand for accurate gesture identification.

In the provided image of the sign alphabet for the deaf (Fig.4), dynamic arrows indicate the movement accompanying the gesture. Analyzing these movements allows for a better understanding of how gestures vary and what additional complexities arise in their recognition. The arrows show the direction of hand or finger movement, which requires the recognition system to account for not only static hand positions but also their dynamics. In some gestures, hand movement is accompanied by a change in its shape, adding a level of complexity for recognition systems, as they must accurately determine both the movement and the final hand configuration.

Despite the presence of dynamic elements, the overall variability of alphabet gestures is lower compared to sign language words. Sign alphabets include a limited number of fixed positions and movements, making them easier to recognize. While sign language words may involve complex sequences of hand movements and configurations, sign letters are usually distinct, easily distinguishable signs. This allows for the use of simpler and more efficient algorithms for their recognition, ensuring high accuracy and reliability of the system.

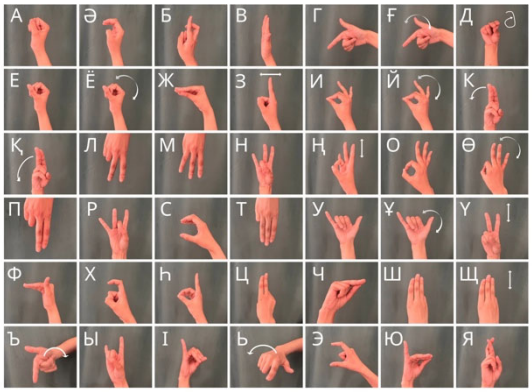


Fig.4.Kazakh dactyl alphabet

Nevertheless, the 2DCNN (Fig.5) architecture remains an important tool in the arsenal of gesture recognition methods, and its further adaptation and combination with other approaches, such as lip reading or hybrid solutions, can improve results for gesture recognition tasks.

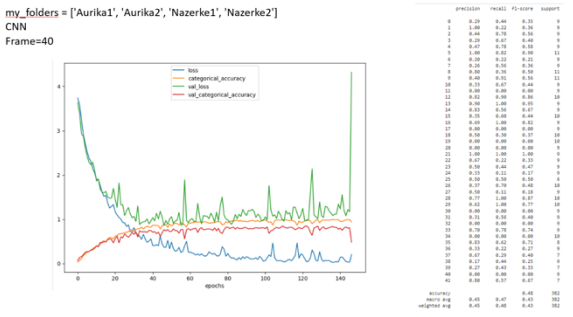


Fig.5. 2DCNN architecture results

Works usingthe LSTM1024 architecture(Fig.6) have shown that this model can effectively work with sequential ortemporal data, considering information about past events to make real-time decisions. The experiments included data collection where each gesture sample was represented by 60 frames, and the model was traine don a large and diverse data set containing 89,700 frames. The LSTM1024 model, trained using the Dropout method and the soft maxactivation function, demonstrated high accuracy

on both training and validation datasets, indicating its capability to recognizege stures even when involving demonstrators of differentages.

The LSTM1024 architecture demonstrated its effectiveness in real-time for recognizing gestures of the Kazakhsign language. The system showed high gesture classification accuracy, as evidenced by high accuracy values on test and validation datasets. The experimental results also showed that the model could accurately recognizegestures despite possible errors by demonstrators of different ages. This indicates there liability and applicability of the proposed model for real usein gesture recognition systems. Moreover, in recognizing the Kazakhdactyl alphabet, the LSTM1024 architecture performed better compared to other models, indicating its capability to accurately recognize both static and dynamic elements of th ealphabet.

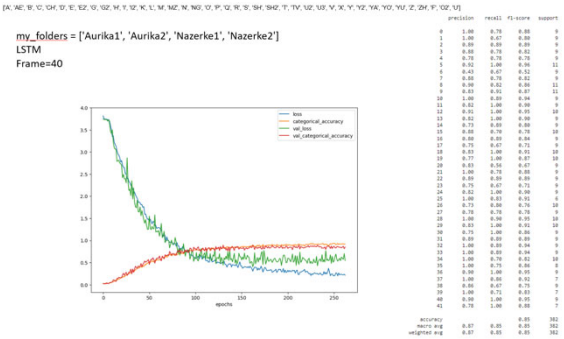


Fig.6. LSTM1024 architecture results

2.3 Additional Features for Recognition

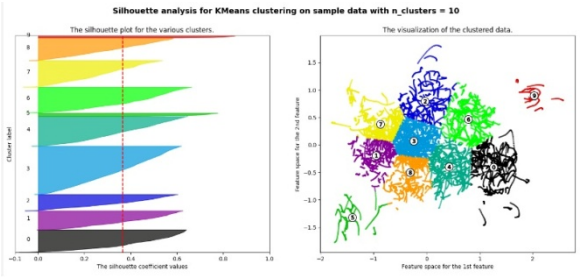


Fig.7. Silhouette analysis for KMeans clustering on sample data with n_clusters = 10

As part of the experiment, the video recordings of the dactyl alphabet were manually divided into classes based on the visual characteristics of lip movements, resulting in the identification of 10 initial sound groups. However, subsequent clustering analysis showed that this classification was overly detailed and did not provide the expected accuracy (Fig.7).

To analyze the lip coordinate data, we used dimensionality reduction methods (PCA, t-SNE, UMAP), various clustering models (K-means, DBSCAN), dynamic time warping (DTW), and tools for clustering time series using tslearn. These methods enabled a comprehensive analysis of the data and the identification of patterns and groupings.

```
0:  'A', 'AE', 'YA'
1:  'B', 'P', 'P2', 'F', 'V', 'M'
2:  'O', 'U', 'U2', 'U3', 'O2', 'YO', 'YU'
3:  'CH', 'E2', 'G2', 'H', 'NG', 'R', 'SH2', 'Y2', 'ZH'
4:  'C', 'D', 'E', 'G', 'I', 'K', 'L', 'N', 'Q', 'S', 'SH', 'T', 'X', 'Y', 'Z'
```

Fig.8. 5 final classes

Using these methods, the 11 initial classes were reduced to five final classes (Fig.8).

Based on the obtained data, these five final classes were classified using machine learning methods: Decision Tree, Random Forest, and XGBoost. Among all the methods, XGBoost demonstrated the highest accuracy and stability, showing the best results on the test dataset.

These results confirm that approaches using dimensionality reduction methods and various clustering models, combined with powerful machine learning algorithms such as XGBoost, are most effective for recognizing lip shapes when demonstrating the alphabet (Fig. 9).

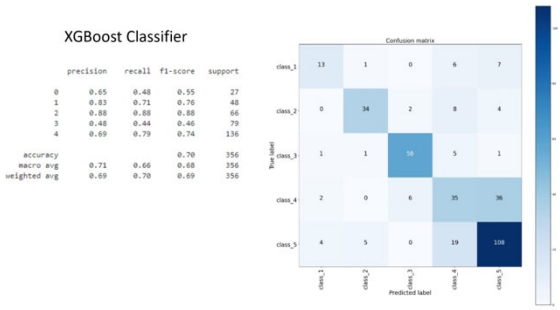


Fig.9. XG Boostclustering

3.3 Analysis of Results

The LSTM1024 model showed a significant reduction in losses on the training dataset, indicating successful training. The model's accuracy on training and validation datasets reached high values, with a macro-average accuracy of 0.87 and a weighted average accuracy of 0.85. The model effectively recognized most letters of the Kazakh dactyl alphabet, making only minor errors in some classes.

The 2DCNN model exhibited unstable losses, especially on the validation dataset. The model's accuracy was significantly lower compared to LSTM1024, with a macro-average accuracy of 0.45 and a weighted average accuracy of 0.45. The 2DCNN model made more errors in recognizing letters, indicating the need for further tuning and improvement of the architecture to achieve better results.

In the experiment, the video recordings of the dactyl alphabet were manually divided into classes based on the visual characteristics of lip movements, resulting

in the identification of 11 initial sound groups. However, subsequent clustering analysis showed that this classification was overly detailed and did not provide the expected accuracy. To analyze the lip coordinate data, we used dimensionality reduction methods (PCA, t-SNE, UMAP), various clustering models (K-means, DBSCAN), dynamic time warping (DTW), and tools for clustering time series using tslearn. These methods enabled a comprehensive analysis of the data and the identification of patterns and groupings.

Using these methods, the 11 initial classes were reduced to five final groups. Based on the obtained data, these five final groups were classified using machine learning methods: Decision Tree, Random Forest, and XGBoost. Among all the methods, XGBoost demonstrated the highest accuracy and stability, showing the best results on the test dataset.

The experiment showed that despite the lower variability of the sign alphabet compared to sign language words, the accuracy of word gesture recognition was higher than that of the alphabet. This result was achieved by comparing 40 variable sign language words and 42 letters of the sign alphabet. This phenomenon may be due to several factors worth considering in more detail.

The sign alphabet consists of fixed positions and movements for each letter, making them predictable and standardized. However, due to the need to accurately recognize each letter, deep learning models may face difficulties in processing fine details and differences in hand configurations, especially under conditions of low image quality or insufficient lighting.

On the other hand, sign language words may include smoother and more natural movements, which can facilitate their recognition. In dynamic gestures used for words, the system can use context and movement sequences to improve recognition accuracy. Moreover, the variability of sign language words may be more expressive, helping the model to better distinguish gestures.

IV. CONCLUSION

The practical application of deep learning requires the use of a single architecture for both words and the alphabet, which is critically important for creating effective sign language recognition systems. We believe that the scientific solution should be effective in terms of practical application, providing high accuracy and reliability.

Our research results confirm that the accuracy of sign language word recognition can be higher than that of the sign alphabet, despite the latter's lower variability. This emphasizes the importance of adapting deep learning models to specific gesture recognition tasks. For the sign alphabet, more complex algorithms and data preprocessing methods may be required to improve recognition accuracy. At the same time, for sign language words, models can use additional contextual cues and movement dynamics to achieve better results.

Using multimodal approaches and adapting models significantly improves gesture recognition quality, ultimately enhancing the lives of deaf and hard-of-hearing people. Integrating the sign alphabet and sign language words into a single recognition system provides a more complete and accurate perception of information, which is especially important for communication and interaction in everyday life. The results of this research can serve as a starting point for creating a more inclusive society, providing deaf and hard-of-hearing people with equal opportunities for communication and self-expression.

ACKNOWLEDGMENTS

This work is part of the AP19175452 project funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan, aimed at developing automated interpretation models for Kazakh sign language. We express our special gratitude to the children of Нұрзада Амангелді: Зейнетулла Амирand Зейнетулла Томирис, who assistedintesting the system and collecting data for training, aswellasinpreviousworksinthisfield. Their contribution and participation were invaluable in achieving the goals of our research.

FUTURE WORK

After identifying the impact of demonstration variability on the application of deep neural networks for recognizing the sign alphabet, as an integral part of sign language recognition, we have begun incorporating lipreading. Including lip reading in gesture recognition models has the potential to further enhance the accuracy and reliability of these systems, especially inconditions of low variability. This line of research will be continued with the aim of creating more versatile and adaptable gesture recognition systems capable of effectively handling various forms and conditions of demonstration.

REFERENCES

[1] Aly, S., & Aly, W. (2020). DeepArSLR: A novel signer-independent deep learning framework for isolated arabic sign language gestures recognition. IEEE Access, 8, 83199–83212. https://doi.org/10.1109/ACCESS.2020.2990699

[2] Kothadiya, D. R., Bhatt, C. M., Rehman, A., Alamri, F. S., & Saba, T. (2023). SignExplainer: An Explainable AI-Enabled Framework for Sign Language Recognition With Ensemble Learning. IEEE Access, 11. https://doi.org/10.1109/ACCESS.2023.3274851

[3] Sahoo, J. P., Prakash, A. J., Pławiak, P., &Samantray, S. (2022). Real-time hand gesture recognition using fine-tuned convolutional neural network. Sensors, 22(3), 706. https://doi.org/10.3390/s22030706

[4] Ewe, E. L. R., Lee, C. P., Kwek, L. C., & Lim, K. M. (2022). Hand Gesture Recognition via Lightweight VGG16 and Ensemble Classifier. Applied Sciences (Switzerland), 12(15). https://doi.org/10.3390/app12157643

[5] Tan, C. K., Lim, K. M., Chang, R. K. Y., Lee, C. P., &Alqahtani, A. (2023). HGR-ViT: Hand Gesture Recognition with Vision Transformer. Sensors, 23(12), 5555. https://doi.org/10.3390/s23125555

[6] Papastratis, I., Dimitropoulos, K., &Daras, P. (2021). Continuous sign language recognition through a context-aware generative adversarial network. Sensors, 21(7). https://doi.org/10.3390/s21072437

[7] Al-Hammadi, M., Bencherif, M. A., Alsulaiman, M., Muhammad, G., Mekhtiche, M. A., Abdul, W., … &Ghaleb, H. (2022). Spatial attention-based 3d graph convolutional neural network for sign language recognition. Sensors, 22(12), 4558. https://doi.org/10.3390/s22124558

[8] Chen, H., Jiang, Y., & Ko, H. (2022). Pose-Guided Graph Convolutional Networks for Skeleton-Based Action Recognition. IEEE Access, 10. https://doi.org/10.1109/ACCESS.2022.3214812

[9] Bird, J. J., Ekárt, A., &Faria, D. R. (2020). British sign language recognition via late fusion of computer vision and leap motion with transfer learning to american sign language. Sensors (Switzerland), 20(18). https://doi.org/10.3390/s20185151

[10] Li, W., Luo, Z., Jin, Y., & Xi, X. (2021). Gesture recognition based on multiscale singular value entropy and deep belief network. Sensors (Switzerland), 21(1). https://doi.org/10.3390/s21010119

[11] Wu, M., Du, D., Li, Y., Bai, W., & Liu, W. (2021). Multi-cascade perceptual human posture recognition enhancement network. IEEE Access, 9, 64256-64266. https://doi.org/10.1109/access.2021.3074541

[12] AmangeldyNurzada, Ukenova Aru, BekmanovaGulmira, RazakhovaBibigul, Milosz Marek, KudubayevaSaule. Continuous Sign Language Recognition and Its Translation into Intonation-Colored Speech. Sensors. Издатель: Multidisciplinary Digital Publishing Institute. ISSN 14248220.DOI: 10.3390/s23146383. Sensors 2022, 23(14), 6383 (Article)

[13] NurzadaAmangeldy, Marek Milosz, SauleKudubayeva, AkmaralKassymova, GulsimKalakova, Lena Zhetkenbay. A Real-Time Dynamic Gesture Variability Recognition Method Based on Convolutional Neural Networks. Applied Sciences. Издатель: Multidisciplinary Digital Publishing Institute. ISSN 2076-3417. DOI:10.3390/app131910799. Appl. Sci. 2023, 13(19), 10799 (Article)

[14] NurzadaAmangeldy, SauleKudubayeva, AkmaralKassymova, ArdakKaripzhanova, BibigulRazakhova, SerikbayKuralov.Sign Language Recognition Method Based on Palm Definition Model and Multiple Classification. Sensors. Издатель: Multidisciplinary Digital Publishing Institute. ISSN 14248220. DOI: 10.3390/s22176621. Sensors 2022, 22(17), 6621. (Article)

[15] AmangeldyNurzada, KudubayevaSaule, RazakhovaBibigul, AsselMukanova, NaziraTursynova. Comparative Analysis Of Classification Methods Of The Dactyl Alphabet Of The Kazakh Language. Journal of Theoretical and Applied Information Technology. Издатель: Little Lion Scientific. ISSN 1992-8645. Volume 100, Issue 19, Pages 5506 - 5513 (Article)

★★★